SOLUTIONS - KERNEL METHODS

STATISTICAL LEARNING COURSE, ECOLE NORMALE SUPÉRIEURE

Raphaël Berthier
raphael.berthier@inria.fr

1. Examples of positive definite kernels

(1) (a) Denote $k = k_1 + k_2$. Let $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in \mathcal{X}$. Then

$$\sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) = \underbrace{\sum_{i,j} \alpha_i \alpha_j k_1(x_i, x_j)}_{\geq 0} + \underbrace{\sum_{i,j} \alpha_i \alpha_j k_2(x_i, x_j)}_{\geq 0} \geq 0.$$

(b) Denote $k = k_1 k_2$. Let $x_1, \ldots, x_n \in \mathcal{X}$ and K_1, K_2 the Gram matrices associated to kernels k_1, k_2 at the points $x_1, \ldots, x_n$. We show that $K = K_1 \odot K_2$, the Gram matrix associated to k, is positive definite. Here, $\odot$ denotes the Hadamard product (i.e., the pointwise product). As K_1 is a symmetric positive semi-definite matrix, one can diagonalize $K_1 = \sum_i \lambda_i u_i u_i^T$. Then

$$K = \sum_i \lambda_i u_i u_i^T \odot K_2$$

But for any vector u :

$$\sum_{ij} \alpha_i \alpha_j (uu^T \odot K_2)_{ij} = \sum_{ij} \alpha_i \alpha_j (K_2)_{ij} u_i u_j = (\alpha \odot u)^T K_2 (\alpha \odot u) \geq 0$$

Thus by summing non-negative terms, $\sum_{i,j} \alpha_i \alpha_j K_{ij} \geq 0$.

(2) Consider $\mathcal{H} = L^2(\mathbb{R})$ and $\phi(x) = \mathbf{1}_{\{t \leq x\}}$.

(3) Similarly,

$$\frac{1}{x+y} = \int_0^1 t^{x-\frac{1}{2}} t^{y-\frac{1}{2}} dt = \langle \phi(x), \phi(y) \rangle_{L_2([0,1])}.$$

Thus $\frac{1}{x+y}$ is a positive definite kernel. Now xy is the standard scalar product on $\mathbb{R}$, and by product k is a positive definite kernel.

(4) Denote n the cardinal of X. For $A \subset X$, denote $\Phi(A)$ the indicator function of A. Then

$$|A \cap B| = \phi(A)^T \phi(B)\,,$$

thus $|A \cap B|$ is a positive definite kernel. Further, denoting A^c the complement of A,

$$
\begin{aligned}
\frac{1}{|A \cup B|} &= \frac{1}{n - |A^c \cap B^c|} \\
&= \frac{1}{n\left(1 - \frac{|A^c \cap B^c|}{n}\right)} \\
&= \frac{1}{n(1 - \frac{\phi(A^c)^T \phi(B^c)}{n})} \\
&= \frac{1}{n} \sum_{i=0}^{\infty} (\frac{\phi(A^c)^T \phi(B^c)}{n})^i
\end{aligned}
$$

which is a positive definite kernel by sum and products of positive definite kernels. Finally, by a final product, K is a positive definite kernel.

(5)

$$
\mathrm{GCD}(n, m) = \prod_{p_i} p_i^{\min(\psi_i(m), \psi_i(n))},
$$

where the p_i are the prime numbers and where $\psi_i(m)$ give the power of p_i in the decomposition of m. We see this as a product of kernels : indeed, consider the feature map

2. Distance in the feature space

(1) (a)

$$
\|\phi(x) - \phi(y)\|^2 = k(x, x) - 2k(x, y) + k(y, y)
$$

(b)

$$
\|\phi(x) - \phi(y)\|^2 = \frac{(x - y)^2}{x + y}.
$$

(2) (a)

$$
\|\phi(x) - \mu_+\|^2 = k(x, x) + \frac{1}{n_+^2} \sum_{i, y_i = 1} \sum_{j, y_j = 1} k(x_i, x_j) - \frac{2}{n_+} \sum_{i, y_i = 1} k(x, x_i)
$$

(b) We output $y = +1$ if

$$
\frac{1}{n_+^2} \sum_{i, y_i = 1} \sum_{j, y_j = 1} k(x_i, x_j) - \frac{2}{n_+} \sum_{i, y_i = 1} k(x, x_i) \leq \frac{1}{n_-^2} \sum_{i, y_i = -1} \sum_{j, y_j = -1} k(x_i, x_j) - \frac{2}{n_-} \sum_{i, y_i = -1} k(x, x_i)
$$

and -1 otherwise.

(c)

$$
\|\phi(x) - \frac{1}{n_+} \sum_{i, y_i = 1} \phi(x_i)\|^2 \leq \|\phi(x) - \frac{1}{n_-} \sum_{i, y_i = -1} \phi(x_i)\|^2 \Leftrightarrow \sum_{i, y_i = 1} k(x, x_i) \geq \sum_{i, y_i = -1} k(x, x_i)
$$

(d) The application is straightforward. However, it sheds light on the importance of the choice of the kernel k. It decides how samples influence the classification rule of other points. The choice of the kernel can thus be seen as a choice of “similarity between points”.